INFORMATION AND CONTROL 50, 110–118 (1981)

Left Universal Context-Free Grammars and Homomorphic Characterizations of Languages

SADAKI HIROSE* AND MASAKAZU NASU

Research Institute of Electrical Communication, Tohoku University, Sendai, Japan

A context-free grammar G with terminal vocabulary Σ is left universal for a class of languages $\mathscr{L}$ with respect to a class of languages $\mathscr{L}_1$ if for each language $L \subset \Sigma^*$ in $\mathscr{L}$, a control language C in $\mathscr{L}_1$ can be found such that G controlled by C generates L by leftmost derivations. We show that for a class of languages $\mathscr{L}$ and a class of languages $\mathscr{L}_1$, if $\mathscr{L}_1$ is closed under homomorphism and inverse homomorphism, then for each alphabet Σ, the following two statements are equivalent. (1) There exists a context-free grammar with terminal vocabulary Σ which is left universal for $\mathscr{L}$ with respect to $\mathscr{L}_1$. (2) There exist a context-free language L_1 and a homomorphism h such that each language $L \subset \Sigma^*$ in $\mathscr{L}$ equals $h(L_1 \cap L')$ for some L' in $\mathscr{L}_1$. We also give some applications of this result.

Kasai (1975) showed that for each alphabet Σ, there exists a context-free grammar G such that for each context-free language L over Σ a regular control language C can be found such that G controlled by C generates L by leftmost derivations. Kasai called such a grammar *universal.* (Greibach (1978) called it *left universal.*) Further studies on "universal grammar" and its extended notions have been done by Hart (1976), Rozenberg (1977), Greibach (1978), and Hirose and Nasu (1980).

In this paper, we consider left universal context-free grammars in the sense of Greibach's definition. Greibach (1978) defined that a grammar G with terminal vocabulary Σ is *left universal for a class of languages $\mathscr{L}$ and Σ with respect to a class of languages $\mathscr{L}_1$* if for each language L over Σ in $\mathscr{L}$, a control language C in $\mathscr{L}_1$ can be found such that G controlled by C generates L by leftmost derivations. We show that for a class of languages $\mathscr{L}$ and a class of languages $\mathscr{L}_1$, if $\mathscr{L}_1$ is closed under homomorphism and inverse homomorphism, then for each alphabet Σ, there exists a left universal context-free grammar for $\mathscr{L}$ and Σ with respect to $\mathscr{L}_1$ if and only if the languages in $\mathscr{L}$ have the same type of homomorphic characterization as the well-known homomorphic characterization of the context-free languages due

* Present address: Software Laboratory, Fujitsu Laboratories Ltd., 1015 Kamikodanaka, Nakahara-ku, Kawasaki, 211, Japan.

0019-9958/81/080110–09$02.00/0

to Chomsky (1962) and Stanley (1965), relative to $\mathscr{L}_1$, that is, there exist a context-free language L_1 and a homomorphism h such that each language L over Σ in $\mathscr{L}$ can be represented as

$$L = h(L_1 \cap L')$$

for some L' in $\mathscr{L}_1$. Furthermore, as applications of this result, we present new homomorphic characterizations similar to the homomorphic characterization of the context-free languages due to Chomsky and Stanley, for the recursively enumerable languages and the linear context-free languages.

For background information on formal languages and grammars, the reader is referred to Salomaa (1973) and Harrison (1978).

Let $G = (V, \Sigma, P, S)$ be a context-free grammar.[1] For any $\pi \in P$, if $\pi = A \to \alpha$ with $A \in V - \Sigma$ and $\alpha \in V^*$, then we write

$$wA\beta \underset{l\ \ G}{\overset{\pi}{\Longrightarrow}} w\alpha\beta$$

for any $w \in \Sigma^*$ and $\beta \in V^*$. Let α, $\beta \in V^*$ and $x = \pi_1 \pi_2 \cdots \pi_n$ with $\pi_i \in P$ for $i = 1,..., n$. We write

$$\alpha \underset{l\ \ G}{\overset{x}{\Longrightarrow}} \beta$$

if there exist α_0, α_1,..., $\alpha_n \in V^*$ such that $\alpha_0 = \alpha$, $\alpha_n = \beta$, and

$$\alpha_{i-1} \underset{l\ \ G}{\overset{\pi_i}{\Longrightarrow}} \alpha_i$$

for $i = 1,..., n$. For a language $C \subset P^+$, we define

$$L_C^l(G) = \left\{ w \in \Sigma^* \mid S \underset{l\ \ G}{\overset{x}{\Longrightarrow}} w \text{ for some } x \in C \right\},$$

and we define $L(G) = L_{P^+}^l(G)$.

DEFINITION. Let Σ be an alphabet and let G be a context-free grammar over Σ.[2] Let $\mathscr{L}$ and $\mathscr{L}_1$ be classes of languages. Then G is *left universal* for $\mathscr{L}$ with respect to $\mathscr{L}_1$ if for any language $L \in \mathscr{L}$ over Σ, there exists a language $C \in \mathscr{L}_1$ such that $L_C^l(G) = L$.

[1] In a context-free grammar $G = (V, \Sigma, P, S)$, $V - \Sigma$ is the set of *nonterminals*, Σ is the set of *terminals*, S in $V - \Sigma$ is the *start symbol*, and P is the set of *productions* of the form $A \to \alpha$, A in $V - \Sigma$ and α in V^*.

[2] A grammar with terminal vocabulary Σ is called a *grammar over Σ*.

THEOREM 1. *Let $\mathscr{L}$ be a class of languages and let $\mathscr{L}_1$ be a class of languages which is closed under homomorphism and inverse homomorphism. Then, for each alphabet Σ, the following two statements are equivalent.*

(1) *There exists a context-free grammar over Σ which is left universal for $\mathscr{L}$ with respect to $\mathscr{L}_1$.*

(2) *There exist an alphabet Σ_1, a context-free language L_1 over Σ_1, and a homomorphism h: $\Sigma_1^* \to \Sigma^*$ such that for any language $L \in \mathscr{L}$ over Σ, there exists a language $L' \in \mathscr{L}_1$ over Σ_1 such that*

$$L = h(L_1 \cap L').$$

Proof. We assume (1). Let $G = (V, \Sigma, P, S)$ be a context-free grammar which is left universal for $\mathscr{L}$ with respect to $\mathscr{L}_1$. We define a context-free grammar

$$G_1 = (V_1, \Sigma_1, P_1, S),$$

where $V_1 = V \cup P$, $\Sigma_1 = \Sigma \cup P$, and P_1 is defined as follows: Corresponding to each production $\pi = A \to \alpha$ in P, the production $\pi' = A \to \pi\alpha$ is in P_1. Let g: $\Sigma_1^* \to P^*$ be the homomorphism defined by

$$g(\pi) = \pi \quad (\pi \in P) \qquad \text{and} \qquad g(a) = \lambda\,[3] \quad (a \in \Sigma),$$

and let h: $\Sigma_1^* \to \Sigma^*$ be the homomorphism defined by

$$h(\pi) = \lambda \quad (\pi \in P) \qquad \text{and} \qquad h(a) = a \quad (a \in \Sigma).$$

Then we can easily show the following:

(i) For any language $C \in \mathscr{L}_1$ over P,

$$L_C^l(G) = h(L(G_1) \cap g^{-1}(C)).$$

To show (i), assume that $w \in L_C^l(G)$. Then there exists $x = \pi_1 \cdots \pi_n$ in C with $\pi_1, \ldots, \pi_n \in P$ such that

$$S \xRightarrow[l\quad G]{x} w.$$

Let $x' = \pi_1' \cdots \pi_n'$ and let w' be the word such that

$$S \xRightarrow[l\quad G_1]{x'} w'.$$

[3] We denote by λ the empty word.

Then since the above derivations are leftmost ones, it follows from the definition of G_1 that w' is of the form

$$w' = \pi_1 w_1 \pi_2 w_2 \cdots \pi_n w_n,$$

where $w_1,\ldots, w_n \in \Sigma^*$ and $w_1 \cdots w_n = w$. We have $w = h(w')$, $w' \in L(G_1)$, and $g(w') = \pi_1 \cdots \pi_n \in C$. Hence

$$w \in h(L(G_1) \cap g^{-1}(C)).$$

Conversely, assume that $w \in h(L(G_1) \cap g^{-1}(C))$. Then there exists $w' \in L(G_1) \cap g^{-1}(C)$ such that $h(w') = w$. Since $w' \in L(G_1)$, there exists $x' \in P_1^+$ such that

$$S \xRightarrow[l \quad G_1]{x'} w'.$$

It follows from the definition of G_1 that there exist $\pi_1,\ldots, \pi_n \in P$ such that $x' = \pi_1' \cdots \pi_n'$, and w' is written in the form

$$w' = \pi_1 w_1 \pi_2 w_2 \cdots \pi_n w_n,$$

where $w_1,\ldots, w_n \in \Sigma^*$. It also follows from the definition of G_1 that

$$S \xRightarrow[l \quad G]{x} w_1 \cdots w_n,$$

where $x = \pi_1 \cdots \pi_n$. Therefore, since $w = h(w') = w_1 \cdots w_n$ and $x = g(w') \in C$, we have $w \in L_C^l(G)$.

Thus we have shown (i).

Since G is left universal for $\mathscr{L}$ with respect to $\mathscr{L}_1$, and $\mathscr{L}_1$ is closed under inverse homomorphism, (2) follows from (i).

Thus we have proved that (1) implies (2).

Assume (2). Let $G_1 = (V_1, \Sigma_1, P_1, S)$ be a context-free grammar in Greibach normal form which generates L_1.[4] Let g: $P_1^* \to \Sigma_1^*$ be the homomorphism defined as follows: For each element $\pi = A \to a\alpha$ of P_1 with $A \in V_1 - \Sigma_1$, $a \in \Sigma_1 \cup \{\lambda\}$, and $\alpha \in (V_1 - \Sigma_1)^*$,

$$g(\pi) = a.$$

Then, clearly the following holds.

[4] A context-free grammar $G = (V, \Sigma, P, S)$ is said to be in *Greibach normal form* if each production is one of the forms $A \to aB_1 \cdots B_n$, $A \to a$, and $S \to \lambda$, where $B_1,\ldots, B_n \in (V - \Sigma) - \{S\}$ and $a \in \Sigma$.

(ii) For $x \in P_1^+$, if

$$S \underset{l \;\; G_1}{\overset{x}{\Longrightarrow}} w, \qquad w \in \Sigma_1^*,$$

then

$$g(x) = w.$$

Define a context-free grammar $G = (V, \Sigma, P, S)$, where $V = (V_1 - \Sigma_1) \cup \Sigma$ and P is defined as follows: Corresponding to each production $\pi = A \to a\alpha$ in P_1 where $A \in V_1 - \Sigma_1$, $a \in \Sigma_1 \cup \{\lambda\}$, and $\alpha \in (V_1 - \Sigma_1)^*$, the production

$$\hat{\pi} = A \to h(a)\alpha$$

is in P. Let H: $P_1^* \to P^*$ be the homomorphism defined by $H(\pi) = \hat{\pi}$ ($\pi \in P_1$). Then clearly the following holds.

(iii) For $x \in P_1^+$, $w_1 \in \Sigma_1^*$, and $w \in \Sigma^*$, if

$$S \underset{l \;\; G_1}{\overset{x}{\Longrightarrow}} w_1 \qquad \text{and} \qquad S \underset{l \;\; G}{\overset{y}{\Longrightarrow}} w,$$

where $y = H(x)$, then $h(w_1) = w$.

Now we shall prove the following.

(iv) For each $L' \in \mathscr{L}_1$ over Σ_1,

$$h(L_1 \cap L') = L^l_{H(g^{-1}(L'))}(G).$$

Assume that $w \in h(L_1 \cap L')$. Then there exists $w_1 \in L_1 \cap L'$ with $h(w_1) = w$. Since $w_1 \in L_1$, there exists $x \in P_1^+$ such that

$$S \underset{l \;\; G_1}{\overset{x}{\Longrightarrow}} w_1.$$

It follows from (ii) that

$$g(x) = w_1,$$

and we have

$$S \underset{l \;\; G}{\overset{y}{\Longrightarrow}} h(w_1) = w,$$

where $y = H(x)$. Since $g(x) = w_1$ and $w_1 \in L'$, we have

$$x \in g^{-1}(w_1) \subset g^{-1}(L'),$$

so that

$$y = H(x) \in H(g^{-1}(L')).$$

Therefore,

$$w \in L^l_{H(g^{-1}(L'))}(G).$$

Conversely, assume that $w \in L^l_{H(g^{-1}(L'))}(G)$. Then there exists $y \in H(g^{-1}(L'))$ such that

$$S \underset{l\ \ G}{\overset{y}{\Longrightarrow}} w.$$

Since $y \in H(g^{-1}(L'))$, there exists

$$x \in g^{-1}(L')$$

such that $H(x) = y$. Let $w_1 \in V_1^*$ with

$$S \underset{l\ \ G_1}{\overset{x}{\Longrightarrow}} w_1.$$

By the definition of H, if $x = \pi_1 \cdots \pi_n$ with $\pi_i \in P_1$, then $y = \hat{\pi}_1 \cdots \hat{\pi}_n$ and for each i, π_i and $\hat{\pi}_i$ are the same if we disregard terminals. Hence $w_1 \in \Sigma_1^*$ so that $w_1 \in L_1$. It follows from (ii) that $g(x) = w_1$. Therefore, $w_1 = g(x) \in g(g^{-1}(L')) = L'$. Hence we have

$$w_1 \in L_1 \cap L'.$$

It follows from (iii) that $h(w_1) = w$. Thus we have

$$w \in h(L_1 \cap L').$$

We have proved (iv).

Since $\mathscr{L}_1$ is closed under homomorphism and inverse homomorphism, it follows from (2) and (iv) that G is a context-free grammar over Σ which is left universal for $\mathscr{L}$ with respect to $\mathscr{L}_1$. Thus (1) holds.

We have proved that (2) implies (1). ∎

Now we shall give some applications of the above theorem.

The following theorem due to Chomsky (1962) and Stanley (1965) is well known.

THEOREM 2 (Chomsky, Stanley). *For each alphabet Σ, there exist a Dyck language over an alphabet Σ' and a homomorphism h: $(\Sigma')^* \to \Sigma^*$ such that* *for each context-free language L over Σ, there exists a regular language R over Σ' such that*

$$L = h(D \cap R).$$

It follows from Theorems 1 and 2 that for each alphabet Σ, there exists a context-free grammar over Σ which is left universal for the class of context-free languages with respect to the class of regular languages. Hence another proof of the result of Kasai (1975) has been presented.

The following result is found in Greibach (1978).

THEOREM 3 (Greibach). *For each alphabet Σ, there exists a context-free grammar over Σ which is left universal for the class of recursively enumerable languages with respect to the class of linear context-free languages.*

Using Theorems 1 and 3, we have a homomorphic characterization of Chomsky–Stanley type for the recursively enumerable languages.

THEOREM 4. *For each alphabet Σ, there exist an alphabet Σ_1, a context-free language L_1 over Σ_1, and a homomorphism h: $\Sigma_1^* \to \Sigma^*$ such that for each recursively enumerable language E over Σ, there exists a linear context-free language L' over Σ_1 such that*

$$E = h(L_1 \cap L').$$

Proof. Since the class of linear context-free languages is closed under homomorphism and inverse homomorphism, the theorem follows from Theorems 1 and 3. ∎

Moreover, a stronger version of the above theorem is obtained.

THEOREM 5. *For each alphabet Σ, there exist a Dyck language over an alphabet Σ' and a homomorphism h: $(\Sigma')^* \to \Sigma$ such that for each recursively enumerable language E over Σ, there exists a linear context-free language L' over Σ' such that*

$$E = h(D \cap L'),$$

and for each context-free language L over Σ, there exists a regular language R over Σ' such that

$$L = h(D \cap R).$$

Proof. It follows from Theorems 1 and 3 that there exist an alphabet Σ_1, a context-free language L_1 over Σ_1, and a homomorphism $h_1: \Sigma_1^* \to \Sigma^*$

such that for each recursively enumerable language E over Σ, there exists a linear context-free language L_2 over Σ_1 such that

$$E = h_1(L_1 \cap L_2).$$

Moreover, by the proof of Theorem 1, we can assume that $\Sigma_1 \supset \Sigma$ and $h_1(a) = a$ for all $a \in \Sigma$.

It follows from Theorem 2 that there exist a Dyck language D over an alphabet Σ' and a homomorphism h_2: $(\Sigma')^* \to \Sigma_1^*$ such that each context-free language over Σ_1 equals $h_2(D \cap R')$ for some regular language R' over Σ'. Therefore, there exists a regular language R' over Σ' such that

$$L_1 = h_2(D \cap R').$$

Let E be any recursively enumerable language over Σ. Then there exists a linear context-free language L_2 over Σ_1 such that

$$E = h_1(L_1 \cap L_2).$$

It follows that

$$E = h_1(h_2(D \cap R') \cap L_2) = h_1(h_2(D \cap R' \cap h_2^{-1}(L_2))).$$

Let $h = h_1 h_2$ and let $L' = R' \cap h_2^{-1}(L_2)$. Then we have

$$E = h(D \cap L').$$

Since the class of linear context-free languages is closed under inverse homomorphism and intersection with a regular language, L' is a linear context-free language.

Let L be any context-free language over Σ. Since $\Sigma_1 \supset \Sigma$, L is a context-free language over Σ_1. Hence there exists a regular language R over Σ' such that

$$L = h_2(D \cap R).$$

Since $h_1(a) = a$ for all $a \in \Sigma$, $L = h_1(L)$. Therefore

$$L = h_1(L) = h_1(h_2(D \cap R)) = h(D \cap R). \quad \blacksquare$$

The following result is found in Greibach (1978).

PROPOSITION 1 (Greibach). *For each alphabet Σ, there exists a linear context-free grammar over Σ which is left universal for the class of linear context-free languages with respect to the class of regular languages.*

We also have a homomorphic characterization of Chomsky–Stanley type for the linear context-free languages.

THEOREM 6. *For each alphabet* Σ, *there exist an alphabet* Σ_1, *a linear context-free language* L_1 *over* Σ_1, *and a homomorphism* $h: \Sigma_1^* \to \Sigma$ *such that for each linear context-free language* L *over* Σ, *there exists a regular language* R *over* Σ_1 *such that*

$$L = h(L_1 \cap R).$$

Proof. See the proof of Theorem 1. In the proof that (1) implies (2), in it, if G is linear context-free, then so is G_1. Thus the result follows from the proof and the above proposition. ∎

We note that every language of the form $h(L_1 \cap R)$, where h is a homomorphism, L_1 a linear context-free language, and R a regular language, is a linear context-free language.

ACKNOWLEDGMENT

The authors would like to thank Professors Namio Honda and Masayuki Kimura for their interest in this work.

RECEIVED: September 27, 1980

REFERENCES

CHOMSKY, N. (1962), Context-free grammars and pushdown storage, *M.I.T. Res. Lab. Electron. Quart. Prog. Report.* **65**, 187–194.

GINSBURG, S. (1966), "The Mathematical Theory of Context-Free Languages," McGraw–Hill, New York.

GREIBACH, S. A. (1978), Comments on universal and left universal grammars, context-sensitive languages, and context-free grammar forms, *Inform. Contr.* **39**, 135–142.

HARRISON, M. A. (1978), "Introduction to Formal Language Theory," Addison–Wesley, Reading, Mass.

HART, J. M. (1976), "Two Extensions of KASAI's Universal Context-Free Grammar," Technical Report No. 22-75, Department of Computer Science, University of Kentucky.

HIROSE, S. AND NASU, M. (1980), Some notes on universal grammars, *Trans. IECE Japan* **63-D**, 287–294 (in Japanese).

HOPCROFT, J. E. AND ULLMAN, J. D. (1979), "Introduction to Automata Theory, Languages, and Computation," Addison–Wesley, Reading, Mass.

KASAI, T. (1975), A universal context-free grammar, *Inform. Contr.* **28**, 30–34.

ROZENBERG, G. (1977), A note on universal grammars, *Inform. Contr.* **34**, 172–175.

SALOMAA, A. (1973), "Formal Languages," Academic Press, New York/London.

STANLEY, R. J. (1965), Finite state representations of context-free languages, *M.I.T. Res. Lab. Electron. Quart. Prog. Report* **76**, 276–279.